Section 8

Données manquantes I : mécanismes, vraisemblance et imputation multiple

Séance du 12 mars 2027 • Lectures : Wu (2010), § 3.1 à 3.5 ; Little et Rubin (2019), ch. 1, 3, 6, 10 ; van Buuren (2018), ch. 2 et 4

Julien Miron – notes rédigées à partir du matériel de Lajmi Lakhal-Chaieb et Alexandre Bureau

Objectifs de la séance

À la fin de cette séance, vous devriez être en mesure de :

1. écrire formellement un problème de données manquantes à l'aide des indicatrices de réponse R et distinguer les mécanismes MCAR, MAR et MNAR ;
2. démontrer que l'analyse des cas complets est sans biais sous MCAR, mais biaisée en général sous MAR et MNAR ;
3. énoncer et démontrer le résultat d'ignorabilité : sous MAR et paramètres distincts, on peut baser l'inférence sur la vraisemblance des données observées ;
4. écrire la vraisemblance des données observées dans le modèle normal bivarié et la maximiser directement ou avec l'algorithme EM ;
5. expliquer pourquoi l'imputation simple sous-estime la variance ; décrire l'imputation multiple propre (MICE, augmentation de données) et les règles de Rubin.

1 Mise en contexte

Presque toutes les études biomédicales comportent des données manquantes : questions laissées sans réponse dans un questionnaire, échec d'un génotypage, échantillon sanguin perdu, participants qui abandonnent une étude longitudinale. La question n'est pas de savoir *s'il* manque des données, mais *pourquoi* elles manquent, et ce que cela implique pour l'objet statistique d'intérêt.

Exemple 1.1 (Densité minérale osseuse et génotype UGT2B17)**.** Giroux et coll. (2012) étudient, chez 1663 femmes ménopausées, la densité minérale osseuse (DMO) en fonction du génotype du gène UGT2B17 et de l'hormonothérapie, en ajustant pour l'âge, le poids, le tabagisme et l'activité physique. La DMO, l'âge et le poids sont toujours observés, mais le tabagisme manque pour 15 % des femmes, l'activité physique pour 2 % et le génotype pour environ 0,2 %. Retirer toutes les femmes

ayant au moins une valeur manquante fait perdre près d'une observation sur six. Plus grave : si les femmes qui ne déclarent pas leur tabagisme diffèrent systématiquement des autres (par exemple parce qu'elles fument davantage), l'échantillon restant n'est plus représentatif. Nous reviendrons sur cet exemple à l'exercice 11.3.

Deux principes guideront les deux séances consacrées aux données manquantes. D'abord, toute méthode repose sur une hypothèse sur le *mécanisme* qui a produit les données manquantes, en général invérifiable à partir des données observées et à justifier par la connaissance du mode de collecte. Ensuite, une méthode valable doit fournir à la fois une estimation convergente *et* une mesure d'incertitude qui reflète la perte d'information.

Cette première séance porte sur les méthodes fondées sur un modèle pour les données (vraisemblance, imputation multiple). La séance 9 traitera des méthodes fondées sur un modèle pour le mécanisme de non-réponse (pondération par probabilité inverse).

2 Notation et mécanismes de données manquantes

2.1 Données complètes, observées et manquantes

On considère n unités indépendantes. Pour l'unité i, les données complètes (celles qu'on aurait voulu recueillir) forment un vecteur $Z_i = (Z_{i1}, \ldots, Z_{ip})^\top$. Par exemple :

- pour un vecteur bivarié, $Z_i = (Y_{i1}, Y_{i2})^\top$ et $p = 2$;
- pour une régression avec k covariables, $Z_i = (Y_i, X_{i1}, \ldots, X_{ik})^\top$ et $p = k + 1$;
- pour une étude longitudinale à T visites, $Z_i = (X_i^\top, Y_{i1}, \ldots, Y_{iT})^\top$.

On définit les *indicatrices de réponse*

$$R_{ij} = \begin{cases} 1 & \text{si } Z_{ij} \text{ est observée,} \\ 0 & \text{si } Z_{ij} \text{ est manquante,} \end{cases} \qquad R_i = (R_{i1}, \ldots, R_{ip})^\top \in \{0,1\}^p.$$

Le vecteur R_i est le *patron* de données manquantes de l'unité i. Pour un patron donné, on partitionne $Z_i = (Z_{\text{obs},i}, Z_{\text{man},i})$ en composantes observées et manquantes ; attention, cette partition dépend de R_i. Ce qu'on observe réellement est le couple $(R_i, Z_{\text{obs},i})$, jamais $Z_{\text{man},i}$.

Le point de vue moderne, dû à Rubin (1976), consiste à traiter R comme une variable aléatoire et à modéliser la loi conjointe de (Z, R), que l'on factorise en général comme

$$p(z, r \mid \theta, \psi) = p(z \mid \theta)\, p(r \mid z, \psi), \tag{1}$$

où θ est le paramètre du modèle des données (souvent l'objet d'intérêt) et ψ celui du *mécanisme de données manquantes* $p(r \mid z, \psi)$.

Un patron est dit *monotone* si les variables peuvent être ordonnées de sorte que, lorsque Z_{ij} manque, Z_{ik} manque aussi pour tout $k > j$. C'est le cas typique de l'attrition dans une étude longitudinale (Wu, tableau 3.1) ; nous y reviendrons à la séance 9.

2.2 Les trois mécanismes

Définition 2.1 (Rubin, 1976 ; Little et Rubin, 2019)**.** Le mécanisme de données manquantes est dit

1. **MCAR** (*missing completely at random*, manquant complètement au hasard) si $\mathbb{P}(R_i = r \mid Z_i) = \mathbb{P}(R_i = r)$ pour tout $r \in \{0,1\}^p$: la non-réponse ne dépend d'aucune donnée, observée ou non ;
2. **MAR** (*missing at random*, manquant au hasard) si $\mathbb{P}(R_i = r \mid Z_i) = \mathbb{P}(R_i = r \mid Z_{\text{obs},i})$ pour tout r : sachant les données observées, la non-réponse ne dépend plus des valeurs manquantes ;
3. **MNAR** (*missing not at random*, manquant de façon non aléatoire) dans les autres cas : $\mathbb{P}(R_i = r \mid Z_i)$ dépend de $Z_{\text{man},i}$, même après conditionnement sur $Z_{\text{obs},i}$.

Remarque 2.2.
1. MCAR ⇒ MAR. Les trois catégories sont donc emboîtées : MCAR ⊂ MAR, et MNAR est le complément de MAR.
2. Dans la définition MAR, la partie observée $Z_{\text{obs},i}$ est celle qui correspond au patron r évalué. L'expression « au hasard » est trompeuse : sous MAR, la non-réponse peut dépendre fortement des données ; elle est « au hasard » seulement *conditionnellement* aux données observées.
3. MCAR peut être partiellement contredit par les données (on compare, par exemple, la loi des variables toujours observées entre les unités complètes et incomplètes ; test de Little, 1988, *JASA*, 83, 1198–1202). En revanche, **on ne peut pas distinguer MAR de MNAR à partir des seules données observées** : pour toute loi MAR qui s'ajuste aux données, il existe des lois MNAR qui produisent exactement la même loi de (R, Z_{obs}) (voir la section 4.3). On rend MAR plus plausible en recueillant des variables *auxiliaires* prédictives de la non-réponse.

Exemple 2.3 (Mécanismes dans le cas bivarié)**.** Soit $Z_i = (Y_{i1}, Y_{i2})^\top$ normal bivarié. On tire au hasard une variable « pilote » $J_i \in \{1,2\}$. Nous utiliserons dans les simulations les trois mécanismes suivants, où z_{iJ} désigne la valeur centrée réduite de la variable pilote et $\text{expit}(u) = 1/(1+e^{-u})$:

— MCAR : l'autre variable, $Y_{i,3-J_i}$, manque avec probabilité constante $\text{expit}(a)$;

— MAR : $Y_{i,3-J_i}$ manque avec probabilité $\text{expit}(a + g\, z_{iJ})$. La probabilité qu'une variable manque dépend de l'autre variable, qui, elle, est observée ;

— MNAR : Y_{iJ_i} *elle-même* manque avec probabilité $\text{expit}(a + g\, z_{iJ})$; les grandes valeurs ont tendance à disparaître.

Dans le cas MAR, on vérifie la définition patron par patron : $\mathbb{P}\{R_i = (1,0)^\top \mid Z_i\} = \frac{1}{2}\,\text{expit}(a + g z_{i1})$ ne dépend que de Y_{i1}, observée dans ce patron ; de même pour $(0,1)^\top$; et $\mathbb{P}\{R_i = (1,1)^\top \mid Z_i\}$ dépend de (Y_{i1}, Y_{i2}), tous deux observés. Ce schéma reprend, en le lissant, celui du fichier de simulation du H26 (où la probabilité valait 0 ou 1 selon un seuil).

3 L'analyse des cas complets

L'*analyse des cas complets* (*listwise deletion*) n'utilise que les unités dont toutes les variables sont observées, c'est-à-dire $R_i = (1, \ldots, 1)^\top$. Une variante, l'*analyse des données disponibles*, utilise pour chaque quantité toutes les unités où les variables nécessaires sont observées (par exemple, la moyenne

de Y_1 calculée sur tous les Y_{i1} observés). C'est la méthode par défaut de la plupart des logiciels (`na.action = na.omit` dans R).

3.1 Cas univarié : MCAR contre MNAR

Soit $Y_1, \ldots, Y_n$ i.i.d. d'espérance μ et de variance finie, et R_i l'indicatrice d'observation de Y_i. Avec $p = 1$, il n'y a pas de partie observée sur laquelle conditionner lorsque Y_i manque : le mécanisme est soit MCAR, soit MNAR. Posons $\pi(y) = \mathbb{P}(R = 1 \mid Y = y)$. L'estimateur des cas complets est

$$\widehat{\mu}_c = \frac{\sum_{i=1}^n R_i Y_i}{\sum_{i=1}^n R_i}.$$

Proposition 3.1. *Si* $\mathrm{E}\{\pi(Y)\} > 0$, *alors* $\widehat{\mu}_c \xrightarrow{\mathbb{P}} \mathrm{E}\{Y\pi(Y)\}/\mathrm{E}\{\pi(Y)\}$. *En particulier,* $\widehat{\mu}_c$ *est convergent pour* μ *sous MCAR, mais pas en général sous MNAR.*

Démonstration. En divisant numérateur et dénominateur par n, la loi des grands nombres donne $n^{-1}\sum R_i Y_i \xrightarrow{\mathbb{P}} \mathrm{E}(RY)$ et $n^{-1}\sum R_i \xrightarrow{\mathbb{P}} \mathrm{E}(R)$, d'où, par le théorème de Slutsky, $\widehat{\mu}_c \xrightarrow{\mathbb{P}} \mathrm{E}(RY)/\mathrm{E}(R)$. Par l'espérance itérée,

$$\mathrm{E}(RY) = \mathrm{E}\{\mathrm{E}(RY \mid Y)\} = \mathrm{E}\{Y\ \mathrm{E}(R \mid Y)\} = \mathrm{E}\{Y\pi(Y)\}, \qquad \mathrm{E}(R) = \mathrm{E}\{\pi(Y)\}.$$

Sous MCAR, $\pi(Y) \equiv \pi$ est constante et la limite vaut $\pi\mu/\pi = \mu$. Sinon,

$$\frac{\mathrm{E}\{Y\pi(Y)\}}{\mathrm{E}\{\pi(Y)\}} - \mu = \frac{\mathrm{Cov}\{Y, \pi(Y)\}}{\mathrm{E}\{\pi(Y)\}},$$

qui est non nul dès que la probabilité d'observation est corrélée à Y (positivement si les grandes valeurs sont plus souvent observées). □

3.2 Une variable toujours observée : MAR

Pour créer un mécanisme MAR, il faut une seconde variable V, toujours observée. Soit $Z_i = (Y_i, V_i)$ et $R_i = (R_{i1}, 1)$; les seuls patrons possibles sont (1,1) et (0,1). Le mécanisme est MAR si

$$\mathbb{P}(R_{i1} = 1 \mid Y_i, V_i) = \pi(V_i)$$

ne dépend que de V_i.

Proposition 3.2. *Sous ce mécanisme MAR,* $\widehat{\mu}_c = \sum R_{i1}Y_i / \sum R_{i1} \xrightarrow{\mathbb{P}} \mathrm{E}\{Y\pi(V)\}/\mathrm{E}\{\pi(V)\}$, *qui diffère de* μ *dès que* $\mathrm{Cov}\{Y, \pi(V)\} \neq 0$.

Démonstration. Comme à la proposition 3.1, $\widehat{\mu}_c \xrightarrow{\mathbb{P}} \mathrm{E}(R_1 Y)/\mathrm{E}(R_1)$. En conditionnant sur (Y, V) et en utilisant MAR,

$$\mathrm{E}(R_1 Y) = \mathrm{E}\{Y\ \mathrm{E}(R_1 \mid Y, V)\} = \mathrm{E}\{Y\pi(V)\}, \qquad \mathrm{E}(R_1) = \mathrm{E}\{\pi(V)\},$$

et l'écart à μ vaut $\mathrm{Cov}\{Y, \pi(V)\}/\mathrm{E}\{\pi(V)\}$. Si Y et V sont associées et que π n'est pas constante, cette covariance est en général non nulle. □

Ainsi, **même sous MAR, l'analyse des cas complets est en général biaisée**. Le biais vient de ce que les unités observées ont une distribution de V différente de celle de la population, et donc une distribution de Y différente. On voit déjà poindre deux remèdes : modéliser la loi de Y sachant V (vraisemblance, imputation ; cette séance), ou repondérer les unités observées par $1/\pi(V)$ (séance 9). En effet, $\mathrm{E}\{R_1 Y/\pi(V)\} = \mathrm{E}\{Y\pi(V)/\pi(V)\} = \mu$.

3.3 Une exception importante : la régression

Proposition 3.3. *Considérons le modèle de régression* $\mathrm{E}(Y \mid X) = m(X;\beta)$ *et supposons que la probabilité qu'une unité soit complète ne dépende que des covariables :* $\mathbb{P}(R = \mathbf{1} \mid Y, X) = \pi(X)$. *Alors l'estimateur des cas complets fondé sur les équations* $\sum_i R_i\, D_i(\beta)\{Y_i - m(X_i;\beta)\} = 0$ *(où* D_i *ne dépend que de* X_i*) est convergent pour* β.

Démonstration. L'équation d'estimation est sans biais :

$$\mathrm{E}[R\, D(\beta)\{Y - m(X;\beta)\}] = \mathrm{E}[\pi(X)\, D(\beta)\ \mathrm{E}\{Y - m(X;\beta) \mid X\}] = 0.$$

On conclut par la théorie des estimateurs M (séance 1). □

Ce mécanisme est MAR si X est toujours observée, et MNAR si la non-réponse dépend d'une covariable elle-même manquante ; dans les deux cas, les cas complets restent convergents pour β, au prix d'une perte de précision. Au contraire, si la non-réponse dépend de Y (même observée !), les cas complets sont biaisés. La morale : le biais de l'analyse des cas complets dépend du lien entre la non-réponse et *la réponse sachant les covariables du modèle.*

4 Vraisemblance des données observées et ignorabilité

4.1 La vraisemblance complète des données observées

Ce qu'on observe pour l'unité i est $(R_i, Z_{\mathrm{obs},i})$. Sous la factorisation (1), sa densité s'obtient en intégrant les composantes manquantes :

$$p(r, Z_{\mathrm{obs}} \mid \theta, \psi) = \int p(Z_{\mathrm{obs}}, Z_{\mathrm{man}} \mid \theta)\, p(r \mid Z_{\mathrm{obs}}, Z_{\mathrm{man}}, \psi)\, \mathrm{d}Z_{\mathrm{man}}. \tag{2}$$

(Pour des variables discrètes, on remplace l'intégrale par une somme.) La vraisemblance « pleine » est le produit de ces contributions sur les n unités. Elle exige un modèle pour le mécanisme de non-réponse, ce qu'on aimerait éviter.

4.2 Le théorème d'ignorabilité

Définition 4.1. Les paramètres θ et ψ sont dits *distincts* si l'espace des paramètres conjoint est le produit cartésien $\Theta \times \Psi$ (approche fréquentiste), ou si les lois a priori sont indépendantes (approche bayésienne). Le mécanisme est *ignorable* s'il est MAR et si θ et ψ sont distincts.

Théorème 4.2 (Rubin, 1976)**.** *Si le mécanisme est MAR, la vraisemblance pleine se factorise :*

$$p(r, Z_{\text{obs}} \mid \theta, \psi) = p(r \mid Z_{\text{obs}}, \psi) \;\times\; p(Z_{\text{obs}} \mid \theta), \qquad p(Z_{\text{obs}} \mid \theta) = \int p(Z_{\text{obs}}, Z_{\text{man}} \mid \theta)\, \mathrm{d}Z_{\text{man}}.$$

Si de plus θ et ψ sont distincts, l'estimateur du maximum de vraisemblance de θ et la loi a posteriori de θ s'obtiennent à partir de la seule vraisemblance des données observées $L(\theta \mid Z_{\text{obs}}) \propto \prod_i p(Z_{\text{obs},i} \mid \theta)$, sans modéliser le mécanisme de non-réponse.

Démonstration. Sous MAR, $p(r \mid Z_{\text{obs}}, Z_{\text{man}}, \psi) = p(r \mid Z_{\text{obs}}, \psi)$ ne dépend pas de Z_{man} et peut sortir de l'intégrale (2) :

$$\begin{aligned} p(r, Z_{\text{obs}} \mid \theta, \psi) &= \int p(Z_{\text{obs}}, Z_{\text{man}} \mid \theta)\, p(r \mid Z_{\text{obs}}, \psi)\, \mathrm{d}Z_{\text{man}} \\ &= p(r \mid Z_{\text{obs}}, \psi) \int p(Z_{\text{obs}}, Z_{\text{man}} \mid \theta)\, \mathrm{d}Z_{\text{man}} = p(r \mid Z_{\text{obs}}, \psi)\, p(Z_{\text{obs}} \mid \theta). \end{aligned}$$

La log-vraisemblance pleine s'écrit donc $\ell(\theta, \psi) = \ell_1(\psi) + \ell_2(\theta)$ avec $\ell_2(\theta) = \sum_i \log p(Z_{\text{obs},i} \mid \theta)$. Si l'espace des paramètres est un produit, maximiser ℓ sur $\Theta \times \Psi$ revient à maximiser séparément ℓ_1 sur Ψ et ℓ_2 sur Θ : l'EMV de θ est celui de ℓ_2. Dans le cadre bayésien, avec $\pi(\theta, \psi) = \pi(\theta)\pi(\psi)$, la loi a posteriori se factorise aussi et $p(\theta \mid r, Z_{\text{obs}}) \propto \pi(\theta)\, p(Z_{\text{obs}} \mid \theta)$. □

Remarque 4.3.

1. Le théorème justifie le principe énoncé dans Wu (§ 3.2) : sous MAR et paramètres distincts, l'inférence peut reposer sur la vraisemblance des données observées. C'est pourquoi les modèles mixtes ajustés par maximum de vraisemblance (séance 3) restent valides sous MAR lorsque des réponses manquent dans une étude longitudinale, pourvu que le modèle soit bien spécifié.
2. Pour l'inférence fréquentiste, il faut utiliser la matrice d'information *observée* (hessienne de ℓ_2) : l'information espérée calculée comme si le patron de non-réponse était fixé n'est valide que sous MCAR (Kenward et Molenberghs, 1998).

4.3 MNAR : non-identifiabilité et analyse de sensibilité

Sous MNAR, la factorisation échoue et on ne peut plus ignorer le mécanisme. Pire, le modèle n'est en général pas identifiable. Dans le cas univarié,

$$p(y) = \mathbb{P}(R = 1)\, p(y \mid R = 1) + \mathbb{P}(R = 0)\, p(y \mid R = 0).$$

Les données identifient $\mathbb{P}(R = 1)$ et $p(y \mid R = 1)$, mais ne contiennent *aucune* information sur $p(y \mid R = 0)$, la loi des valeurs non observées : n'importe quel choix de cette loi est compatible avec les données. Toute estimation de $\mathrm{E}(Y)$ repose alors sur une hypothèse non vérifiable. On procède donc à une *analyse de sensibilité* : on postule plusieurs modèles plausibles pour le mécanisme et on examine comment les conclusions varient (Wu, § 3.2 ; Little et Rubin, 2019, ch. 15). Deux grandes familles de modèles sont utilisées.

— **Modèles de sélection** : on suit la factorisation (1), $p(z \mid \theta)\, p(r \mid z, \psi)$, par exemple $\operatorname{logit} \mathbb{P}(R = 1 \mid y) = \psi_0 + \psi_1 y$ (modèle de Heckman en économétrie, modèle de Diggle et Kenward, 1994, en longitudinal). Le paramètre ψ_1 n'est identifié que par la forme paramétrique supposée pour $p(z \mid \theta)$ (par exemple la normalité), ce qui rend l'estimation très fragile ; on fixe plutôt ψ_1 à une grille de valeurs.

— **Modèles de mélange de patrons** (*pattern-mixture*) : on factorise à l'inverse, $p(z, r) = p(r)\, p(z \mid r)$, et on relie la loi des non-répondants à celle des répondants par un paramètre de sensibilité non identifiable, par exemple $p(y \mid R = 0) = p(y - \delta \mid R = 1)$ (« ajustement δ »). L'imputation multiple se prête bien à cette approche : on impute sous MAR, puis on ajoute δ aux valeurs imputées (argument `post` de `mice`) pour une grille de δ.

5 L'exemple normal bivarié

Soit $Z_i = Y_i = (Y_{i1}, Y_{i2})^\top \overset{\text{i.i.d.}}{\sim} \mathcal{N}_2(\mu, \Sigma)$ avec $\mu = (\mu_1, \mu_2)^\top$ et

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}, \qquad \alpha = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)^\top.$$

En données complètes, les EMV sont les moyennes, variances (diviseur n) et la corrélation empiriques. Les observations où $(R_{i1}, R_{i2}) = (0,0)$ ne contiennent aucune information sur α ; on les exclut d'emblée, ce qui revient à les supposer MCAR.

5.1 Données disponibles

L'analyse des données disponibles (appelée « analyse de données complètes » dans le document du H26) donne

$$\widehat{\mu}_j = \frac{\sum_i R_{ij} Y_{ij}}{\sum_i R_{ij}}, \quad \widehat{\sigma}_j^2 = \frac{\sum_i R_{ij}(Y_{ij} - \widehat{\mu}_j)^2}{\sum_i R_{ij}}, \quad \widehat{\rho} = \frac{\sum_i R_{i1} R_{i2} (Y_{i1} - \widehat{\mu}_1)(Y_{i2} - \widehat{\mu}_2)}{\widehat{\sigma}_1 \widehat{\sigma}_2 \sum_i R_{i1} R_{i2}}.$$

Par la section 3, ces estimateurs sont convergents sous MCAR seulement.

5.2 Vraisemblance des données observées

Par le théorème 4.2, sous MAR, on maximise la log-vraisemblance des données observées. La contribution de l'unité i est la densité de sa partie observée : bivariée si les deux composantes sont observées, marginale univariée sinon (la marginale d'une normale bivariée est normale). Ainsi

$$\ell(\alpha) = \ell_{11}(\alpha) + \ell_{10}(\alpha) + \ell_{01}(\alpha), \tag{3}$$

$$\begin{aligned}
\ell_{11}(\alpha) &= \sum_{i=1}^{n} R_{i1} R_{i2} \, \log f_2(Y_{i1}, Y_{i2}; \mu, \Sigma), \\
\ell_{10}(\alpha) &= \sum_{i=1}^{n} R_{i1}(1 - R_{i2}) \, \log f_1(Y_{i1}; \mu_1, \sigma_1^2), \\
\ell_{01}(\alpha) &= \sum_{i=1}^{n} (1 - R_{i1}) R_{i2} \, \log f_1(Y_{i2}; \mu_2, \sigma_2^2),
\end{aligned}$$

où f_1 et f_2 sont les densités normales univariée et bivariée. Les unités incomplètes contribuent à l'estimation de ρ indirectement : dans le patron (1,0), l'information sur μ_1 et σ_1^2 se propage à μ_2 par la corrélation. Il n'y a pas de solution explicite en général (sauf patron monotone) ; on maximise numériquement (Newton-Raphson, quasi-Newton) ou par l'algorithme EM (section 6).

5.3 Imputation simple par régression et sous-estimation de la variance

L'imputation remplace chaque valeur manquante par une valeur plausible, puis applique l'analyse prévue pour les données complètes. Dans le cas bivarié, il faut un modèle pour Y_2 sachant Y_1 (patron (1,0)) et pour Y_1 sachant Y_2 (patron (0,1)). Le plus naturel est la régression linéaire, cohérente avec la normalité :

$$Y_{i2} = \beta_{02} + \beta_{12} Y_{i1} + \varepsilon_{i2}, \quad \varepsilon_{i2} \sim \mathcal{N}(0, \tau_2^2); \qquad Y_{i1} = \beta_{01} + \beta_{11} Y_{i2} + \varepsilon_{i1}, \quad \varepsilon_{i1} \sim \mathcal{N}(0, \tau_1^2).$$

L'*imputation simple par régression* procède ainsi :

1. estimer les coefficients avec les unités complètes, $(R_{i1}, R_{i2}) = (1,1)$;
2. si $(R_{i1}, R_{i2}) = (1,0)$, imputer $\tilde{Y}_{i2} = \widehat{\beta}_{02} + \widehat{\beta}_{12} Y_{i1}$;
3. si $(R_{i1}, R_{i2}) = (0,1)$, imputer $\tilde{Y}_{i1} = \widehat{\beta}_{01} + \widehat{\beta}_{11} Y_{i2}$;
4. analyser le jeu complété comme s'il était complet.

Cette méthode ignore deux sources de variabilité : le terme d'erreur ε (on impute une moyenne conditionnelle, pas une valeur plausible) et l'incertitude sur $(\widehat{\beta}, \widehat{\tau}^2)$. Elle traite en outre les valeurs imputées comme si elles avaient été mesurées. Le calcul suivant quantifie le problème.

Proposition 5.1. *Supposons Y_1 toujours observée, Y_2 manquante pour une fraction q des unités selon un mécanisme MAR (dépendant de Y_1) ou MCAR, et la régression de Y_2 sur Y_1 estimée sur les unités complètes. Alors :*

1. *la moyenne du jeu complété, $\tilde{\mu}_2 = \widehat{\beta}_{02} + \widehat{\beta}_{12} \bar{Y}_1$, est convergente pour μ_2 (c'est l'EMV) ;*
2. *sous MCAR, la variance empirique du jeu complété converge vers $\sigma_2^2\{1 - q(1-\rho^2)\} < \sigma_2^2$;*
3. *sous MCAR, l'erreur type naïve $\tilde{s}_2/\sqrt{n}$ sous-estime l'erreur type asymptotique de $\tilde{\mu}_2$, qui vaut $\sigma_2\{\rho^2 + (1-\rho^2)/(1-q)\}^{1/2}/\sqrt{n}$.*

Démonstration. (1) Les MCO sur les unités complètes donnent $\sum_{\text{obs}} \{Y_{i2} - \widehat{\beta}_{02} - \widehat{\beta}_{12} Y_{i1}\} = 0$, donc la somme des Y_{i2} observées égale la somme des valeurs prédites. La moyenne du jeu complété est donc $n^{-1} \sum_{i=1}^{n} (\widehat{\beta}_{02} + \widehat{\beta}_{12} Y_{i1}) = \widehat{\beta}_{02} + \widehat{\beta}_{12} \bar{Y}_1$. Sous MAR (non-réponse fonction de Y_1), la proposition 3.3 assure la convergence de $(\widehat{\beta}_{02}, \widehat{\beta}_{12})$ vers $(\mu_2 - \beta\mu_1, \beta)$ avec $\beta = \rho\sigma_2/\sigma_1$; $\bar{Y}_1 \xrightarrow{\mathbb{P}} \mu_1$ conclut.

(2) Sous MCAR, la fraction $1-q$ des valeurs complétées a la loi de Y_2, de variance σ_2^2, et la fraction q est égale (à la limite) à $\mu_2 + \beta(Y_1 - \mu_1)$, de même moyenne et de variance $\beta^2\sigma_1^2 = \rho^2\sigma_2^2$. La variance du mélange tend vers $(1-q)\sigma_2^2 + q\rho^2\sigma_2^2$.

(3) Écrivons $\tilde{\mu}_2 = \bar{Y}_{2,c} + \widehat{\beta}_{12}(\bar{Y}_1 - \bar{Y}_{1,c})$, où l'indice c désigne les moyennes sur les $n(1-q)$ unités complètes. Une linéarisation (ou Little et Rubin, 2019, § 7.2) donne $n \operatorname{Var}(\tilde{\mu}_2) \to \sigma_2^2\rho^2 + \sigma_2^2(1-\rho^2)/(1-q)$: le premier terme est la variance de $\beta\bar{Y}_1$, le second celle des résidus, estimés avec seulement $n(1-q)$ unités. Puisque $\rho^2 + (1-\rho^2)/(1-q) \geq 1 \geq 1 - q(1-\rho^2)$, l'erreur type naïve est trop petite, d'autant plus que q est grand et ρ faible. □

L'imputation simple par régression donne donc une moyenne raisonnable, mais des variances et des corrélations biaisées (la corrélation est surestimée, puisque les points imputés tombent exactement sur la droite) et des intervalles de confiance trop étroits. On pourrait corriger la variance analytiquement

(Wu, § 3.3.2), mais c'est peu pratique en général ; l'imputation multiple résout le problème de façon générique. Wu (§ 3.3.1) passe en revue d'autres méthodes naïves (imputation par la moyenne, dernière observation reportée, indicatrice de valeur manquante) qui partagent ce défaut, et sont en plus biaisées sous MAR.

5.4 Illustration en R

Le code complet (fonctions `generer`, `nll`, `em`, `imput.reg`, etc.) est distribué avec les notes ; on montre ici les étapes principales sur un jeu de $n = 200$ observations généré sous le mécanisme MAR de l'exemple 2.3 ($\mu = (1{,}2)$, $\sigma = (0{,}5;\ 1)$, $\rho = 0{,}5$, $a = -0{,}6$, $g = 2$).

```
nll <- function(th, Y) {   # moins la log-vraisemblance des données observées
  m <- th[1:2]; sd <- th[3:4]; r <- th[5]
  if (any(sd <= 0) || abs(r) >= 1) return(1e10)
  S <- diag(sd) %*% matrix(c(1, r, r, 1), 2) %*% diag(sd)
  o1 <- !is.na(Y[, 1]); o2 <- !is.na(Y[, 2])
  -(sum(dmvnorm(Y[o1 & o2, , drop = FALSE], m, S, log = TRUE)) +   # l11
      sum(dnorm(Y[o1 & !o2, 1], m[1], sd[1], log = TRUE)) +        # l10
      sum(dnorm(Y[!o1 & o2, 2], m[2], sd[2], log = TRUE)))         # l01
}
set.seed(7360); Y <- generer(200, "MAR")
table(R1 = 1 * !is.na(Y[, 1]), R2 = 1 * !is.na(Y[, 2]))
round(cc(Y), 3)                                                    # données disponibles
op <- optim(cc(Y), nll, Y = Y, method = "BFGS", hessian = TRUE)
round(op$par, 3); round(sqrt(diag(solve(op$hessian))), 3)
Yi <- imput.reg(Y); round(cc(Yi), 3); sd(Yi[, 2]) / sqrt(200)
```

```
   R2
R1    0   1
  0   0  29
  1  35 136
  mu1   mu2    s1    s2   rho        # données disponibles
0.885 1.850 0.512 0.892 0.317
  mu1   mu2    s1    s2   rho        # EMV
0.906 1.919 0.511 0.899 0.388
0.039 0.071 0.028 0.051 0.078        # erreurs types (hessienne observée)
  mu1   mu2    s1    s2   rho        # imputation simple par régression
0.884 1.893 0.476 0.821 0.378
ET naive de mu2 : 0.0581
```

Sous MAR, les grandes valeurs d'une variable entraînent la disparition de l'autre ; comme $\rho > 0$, les valeurs observées de chaque variable sont décalées vers le bas, ce qu'on voit sur les moyennes des données disponibles. L'EMV corrige ce décalage. L'imputation simple sous-estime nettement les écarts types (0,476 et 0,821) et son erreur type naïve pour μ_2 (0,058) est bien inférieure à celle de l'EMV (0,071). La figure 1 montre pourquoi : les valeurs imputées s'alignent sur la droite de régression.

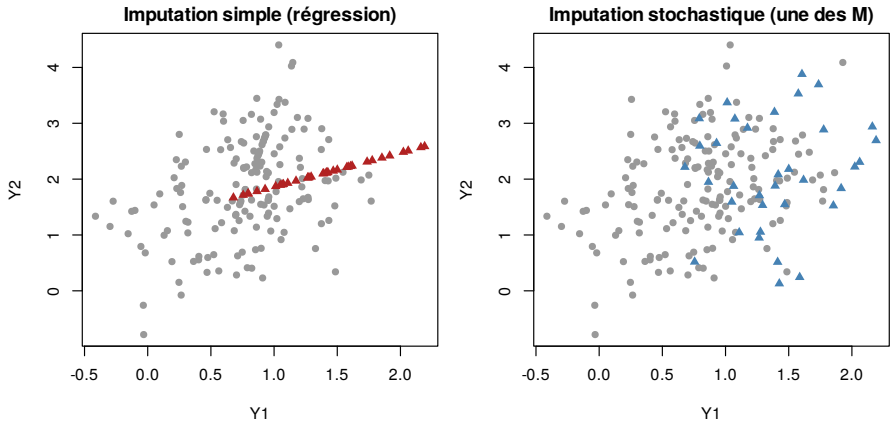


FIGURE 1 – Valeurs de Y_2 imputées (triangles) lorsque seule Y_1 est observée. À gauche : imputation simple par la moyenne conditionnelle, sans dispersion. À droite : un des M jeux imputés par équations chaînées, où l'on ajoute un bruit tiré de la loi prédictive.

6 L'algorithme EM

6.1 Principe

L'algorithme EM (Dempster, Laird et Rubin, 1977 ; Wu, § 3.4) maximise la vraisemblance des données observées $\ell(\theta) = \log p(Z_{\text{obs}} \mid \theta)$ lorsque la vraisemblance des données *complètes* $\ell_c(\theta) = \log p(Z_{\text{obs}}, Z_{\text{man}} \mid \theta)$ est simple à maximiser. Partant de $\theta^{(0)}$, on itère :

Étape E calculer $Q(\theta \mid \theta^{(t)}) = \mathrm{E}\{\ell_c(\theta) \mid Z_{\text{obs}}; \theta^{(t)}\}$, l'espérance de la log-vraisemblance complète sous la loi conditionnelle des données manquantes évaluée en $\theta^{(t)}$;

Étape M poser $\theta^{(t+1)} = \arg\max_\theta Q(\theta \mid \theta^{(t)})$.

Pour une famille exponentielle, l'étape E se réduit au calcul des espérances conditionnelles des statistiques exhaustives, et l'étape M applique les formules des EMV en données complètes à ces statistiques « complétées ».

Proposition 6.1 (Monotonie). *$\ell(\theta^{(t+1)}) \geq \ell(\theta^{(t)})$ pour tout t.*

Démonstration. Comme $p(Z_{\text{obs}}, Z_{\text{man}} \mid \theta) = p(Z_{\text{obs}} \mid \theta)\, p(Z_{\text{man}} \mid Z_{\text{obs}}, \theta)$, on a $\ell(\theta) = \ell_c(\theta) - \log p(Z_{\text{man}} \mid Z_{\text{obs}}, \theta)$. En prenant l'espérance conditionnelle sous $\theta^{(t)}$ des deux côtés (le membre de gauche ne dépend pas de Z_{man}),

$$\ell(\theta) = Q(\theta \mid \theta^{(t)}) - H(\theta \mid \theta^{(t)}), \qquad H(\theta \mid \theta^{(t)}) = \mathrm{E}\{\log p(Z_{\text{man}} \mid Z_{\text{obs}}, \theta) \mid Z_{\text{obs}}; \theta^{(t)}\}.$$

Par l'inégalité de Jensen (ou la positivité de la divergence de Kullback-Leibler), $H(\theta \mid \theta^{(t)}) \leq H(\theta^{(t)} \mid \theta^{(t)})$ pour tout θ. Donc $\ell(\theta^{(t+1)}) - \ell(\theta^{(t)}) = \{Q(\theta^{(t+1)} \mid \theta^{(t)}) - Q(\theta^{(t)} \mid \theta^{(t)})\} - \{H(\theta^{(t+1)} \mid \theta^{(t)}) - H(\theta^{(t)} \mid \theta^{(t)})\} \geq 0$, le premier terme étant positif par définition de l'étape M. □

Sous des conditions de régularité, la suite converge vers un point stationnaire de ℓ (en général un maximum local). La convergence est linéaire, avec un taux égal à la plus grande fraction d'information manquante : EM est lent quand beaucoup d'information manque. EM ne fournit pas directement

d'erreurs types ; on les obtient par la hessienne numérique de ℓ au point de convergence, par la formule de Louis (1982) ou par le bootstrap.

6.2 Illustration : normale bivariée

Les statistiques exhaustives des données complètes sont $\sum_i Y_{ij}$ et $\sum_i Y_{ij}Y_{ik}$, $j,k \in \{1,2\}$. Posons, à l'itération t, $\beta_{2\cdot1} = \sigma_{12}/\sigma_{11}$ et $\sigma_{22\cdot1} = \sigma_{22} - \sigma_{12}^2/\sigma_{11}$ (paramètres de la régression de Y_2 sur Y_1 implicites dans $\Sigma^{(t)}$).

Étape E Pour une unité du patron (1,0),

$$\mathrm{E}(Y_{i2} \mid Y_{i1}) = \mu_2 + \beta_{2\cdot1}(Y_{i1} - \mu_1) =: \tilde{Y}_{i2}, \quad \mathrm{E}(Y_{i2}^2 \mid Y_{i1}) = \tilde{Y}_{i2}^2 + \sigma_{22\cdot1}, \quad \mathrm{E}(Y_{i1}Y_{i2} \mid Y_{i1}) = Y_{i1}\tilde{Y}_{i2};$$

symétriquement pour le patron (0,1). Les unités complètes contribuent leurs valeurs observées.

Étape M

$$\mu^{(t+1)} = \frac{1}{n}\sum_i \tilde{Y}_i, \qquad \Sigma^{(t+1)} = \frac{1}{n}\sum_i (\tilde{Y}_i - \mu^{(t+1)})(\tilde{Y}_i - \mu^{(t+1)})^\top + \frac{1}{n}C,$$

où $\tilde{Y}_i$ est le vecteur complété par les espérances conditionnelles et $C = \mathrm{diag}(n_{01}\sigma_{11\cdot2},\ n_{10}\sigma_{22\cdot1})$ corrige la variance : n_{10} et n_{01} sont les effectifs des patrons (1,0) et (0,1).

L'étape E ressemble à l'imputation simple par régression, mais la régression est réestimée à chaque itération avec toutes les unités et le terme C restitue la variance résiduelle oubliée par l'imputation simple.

```
> e <- em(Y, trace = TRUE)
it  1  mu2 = 1.84956  log-vrais = -335.794401
it  2  mu2 = 1.88183  log-vrais = -334.016739
it 10  mu2 = 1.91894  log-vrais = -333.414144
it 23  mu2 = 1.91903  log-vrais = -333.414141      (convergence)
   Y1    Y2    Y1    Y2
0.906 1.919 0.511 0.899 0.388
```

La log-vraisemblance croît à chaque itération (proposition 6.1) et EM retrouve l'EMV obtenu par `optim`.

Niveau doctorat : vitesse de convergence et information manquante

Près de l'EMV $\widehat{\theta}$, l'application $M : \theta^{(t)} \mapsto \theta^{(t+1)}$ vérifie $\theta^{(t+1)} - \widehat{\theta} \approx DM\,(\theta^{(t)} - \widehat{\theta})$ avec $DM = I_c^{-1} I_m$, où $I_c = -\mathrm{E}\{\partial^2 \ell_c / \partial\theta\partial\theta^\top \mid Z_{\mathrm{obs}}\}$ est l'information complète « espérée conditionnellement » et $I_m = I_c - I_{\mathrm{obs}}$ l'information manquante (principe d'information manquante : $I_{\mathrm{obs}} = I_c - I_m$). Le taux de convergence est la plus grande valeur propre de DM, la plus grande fraction d'information manquante. La méthode SEM (Meng et Rubin, 1991) exploite cette identité pour obtenir $I_{\mathrm{obs}}^{-1} = (I - DM)^{-1} I_c^{-1}$ à partir des itérations d'EM. Cette même fraction d'information manquante réapparaîtra dans les règles de Rubin (section 8.5).

7 Covariables manquantes

Quand seules des *réponses* manquent dans un modèle de régression (données longitudinales débalancées, par exemple), la vraisemblance des données observées est simplement celle du modèle évaluée

sur les réponses observées : on conditionne sur les covariables, toujours connues, et l'inférence par maximum de vraisemblance s'effectue sans modification (Wu, § 3.3.2).

La situation change quand des *covariables* manquent (Wu, p. 101). Écrivons $X = (X_{\text{obs}}, X_{\text{man}})$. Sous MAR, la contribution d'une unité est

$$p(y, x_{\text{obs}} \mid \beta, \alpha) = \int p(y \mid x_{\text{obs}}, x_{\text{man}}; \beta)\, p(x_{\text{obs}}, x_{\text{man}}; \alpha)\, \mathrm{d}x_{\text{man}}.$$

On ne peut plus conditionner sur X : il faut **modéliser la loi des covariables** $p(x; \alpha)$, ou au moins $p(x_{\text{man}} \mid x_{\text{obs}}; \alpha)$, ce qui introduit des paramètres de nuisance et un risque de mauvaise spécification. Avec plusieurs covariables de types mixtes (continues, binaires, ordinales), la spécification d'une loi conjointe devient délicate ; on factorise souvent $p(x_1, \ldots, x_k) = p(x_1)\, p(x_2 \mid x_1) \cdots p(x_k \mid x_1, \ldots, x_{k-1})$, chaque facteur étant un modèle de régression approprié. L'imputation multiple par équations chaînées, présentée plus bas, contourne en partie cette difficulté en spécifiant seulement des lois conditionnelles.

Attention

Dans un modèle d'imputation d'une covariable, il faut inclure la **réponse** Y comme prédicteur. Imputer X à partir des autres covariables seulement revient à imputer sous l'hypothèse d'absence d'association entre X et Y, ce qui atténue l'effet estimé de X vers zéro.

8 Imputation multiple

8.1 Principe

L'imputation multiple (Rubin, 1987 ; Wu, § 3.3.2 et 3.5 ; van Buuren, 2018) procède en trois étapes :

1. **Imputation** : créer M jeux de données complétés, en remplaçant les valeurs manquantes par des tirages aléatoires de leur loi prédictive sachant les données observées ;
2. **Analyse** : appliquer à chaque jeu complété l'analyse prévue pour les données complètes, ce qui donne $\widehat{\theta}^{(m)}$ et une estimation de sa variance $\widehat{V}^{(m)}$, $m = 1, \ldots, M$;
3. **Combinaison** : combiner ces M résultats par les règles de Rubin.

Parmi ses avantages (Wu, p. 106) : une fois les imputations faites, on utilise les méthodes et logiciels habituels ; le modèle d'imputation peut inclure des variables auxiliaires absentes du modèle d'analyse ; la variabilité due aux données manquantes est prise en compte ; et quelques imputations suffisent souvent.

8.2 Justification bayésienne et imputation « propre »

Soit θ la quantité d'intérêt. La loi a posteriori sachant les données observées est un mélange :

$$p(\theta \mid Z_{\text{obs}}) = \int p(\theta \mid Z_{\text{obs}}, Z_{\text{man}})\, p(Z_{\text{man}} \mid Z_{\text{obs}})\, \mathrm{d}Z_{\text{man}}.$$

Si on tire $Z_{\text{man}}^{(m)} \sim p(Z_{\text{man}} \mid Z_{\text{obs}})$, la loi a posteriori est approchée par la moyenne des lois a posteriori en données complétées. Par les formules de l'espérance et de la variance totales,

$$\mathrm{E}(\theta \mid Z_{\text{obs}}) = \mathrm{E}\{\mathrm{E}(\theta \mid Z_{\text{obs}}, Z_{\text{man}}) \mid Z_{\text{obs}}\} \approx \frac{1}{M}\sum_m \widehat{\theta}^{(m)}, \tag{4}$$

$$\mathrm{Var}(\theta \mid Z_{\text{obs}}) = \underbrace{\mathrm{E}\{\mathrm{Var}(\theta \mid Z_{\text{obs}}, Z_{\text{man}}) \mid Z_{\text{obs}}\}}_{\approx \text{ variance intra}} + \underbrace{\mathrm{Var}\{\mathrm{E}(\theta \mid Z_{\text{obs}}, Z_{\text{man}}) \mid Z_{\text{obs}}\}}_{\approx \text{ variance inter}}. \tag{5}$$

Le point crucial est la loi prédictive : avec ψ les paramètres du modèle d'imputation,

$$p(Z_{\text{man}} \mid Z_{\text{obs}}) = \int p(Z_{\text{man}} \mid Z_{\text{obs}}, \psi)\, p(\psi \mid Z_{\text{obs}})\, \mathrm{d}\psi.$$

Pour tirer de cette loi, il faut **d'abord tirer $\psi^{(m)}$ de sa loi a posteriori**, puis $Z_{\text{man}}^{(m)} \sim p(Z_{\text{man}} \mid Z_{\text{obs}}, \psi^{(m)})$. Une imputation qui fixe $\psi = \widehat{\psi}$ (même en ajoutant un bruit résiduel) ignore l'incertitude sur ψ et sous-estime la variance inter-imputations : elle est dite *impropre*. Une imputation qui propage cette incertitude est dite *propre* (Rubin, 1987). Ainsi, dans le cas bivarié, une imputation propre de Y_2 sachant Y_1 tire successivement τ_2^{*2}, puis $(\beta_{02}^*, \beta_{12}^*)$, puis ε^*.

8.3 Imputation par équations chaînées (MICE)

Tirage a posteriori dans une régression linéaire. Soit la régression $y = X\beta + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \tau^2 I)$, ajustée sur n_{obs} unités où y est observée, avec k coefficients. Sous la loi a priori non informative $p(\beta, \tau^2) \propto \tau^{-2}$, la loi a posteriori est

$$\tau^2 \mid y \sim \frac{(n_{\text{obs}} - k)\,\widehat{\tau}^2}{\chi^2_{n_{\text{obs}}-k}}, \qquad \beta \mid \tau^2, y \sim \mathcal{N}_k\big(\widehat{\beta},\ \tau^2 (X^\top X)^{-1}\big),$$

où $\widehat{\beta}$ et $\widehat{\tau}^2 = \mathrm{SCR}/(n_{\text{obs}} - k)$ sont les estimateurs des moindres carrés. C'est la méthode `"norm"` de `mice`.

Algorithme bivarié (document du H26). Notons Y_{ij}^* la valeur courante (imputée) de Y_{ij} lorsqu'elle manque, et $n_j = \sum_i R_{ij}$.

1. Initialiser les valeurs manquantes, par exemple par des tirages au hasard parmi les valeurs observées de la même variable (ou par $\bar{Y}_j$).
2. Répéter les étapes suivantes (un *cycle*) un nombre suffisant de fois :
   (a) avec les unités où $R_{i2} = 1$, régresser Y_2 sur Y_1 (valeurs courantes) : $\widehat{\beta}_{02}, \widehat{\beta}_{12}, \widehat{\tau}_2^2$;
   (b) tirer $\tau_2^{*2} = (n_2 - 2)\widehat{\tau}_2^2 / c$ avec $c \sim \chi^2_{n_2-2}$;
   (c) tirer
   $$\begin{pmatrix}\beta_{02}^* \\ \beta_{12}^*\end{pmatrix} \sim \mathcal{N}_2\left(\begin{pmatrix}\widehat{\beta}_{02} \\ \widehat{\beta}_{12}\end{pmatrix},\ \tau_2^{*2}\left\{\sum_{i=1}^{n} R_{i2}\begin{pmatrix}1 & Y_{i1} \\ Y_{i1} & Y_{i1}^2\end{pmatrix}\right\}^{-1}\right);$$
   (d) pour chaque i du patron (1,0), tirer $\varepsilon_{i2}^* \sim \mathcal{N}(0, \tau_2^{*2})$ et poser $Y_{i2}^* = \beta_{02}^* + \beta_{12}^* Y_{i1} + \varepsilon_{i2}^*$;
   (e)–(h) faire de même pour Y_1 sachant Y_2, avec les unités où $R_{i1} = 1$, et imputer le patron (0,1).

3. Après les cycles, conserver le jeu complété ; recommencer les étapes 1–2 de façon indépendante pour obtenir M jeux.

Cas général. Avec p variables, on spécifie pour chaque variable incomplète Z_j un modèle conditionnel $p(Z_j \mid Z_{-j}, \psi_j)$ adapté à sa nature : régression linéaire (`"norm"`), appariement sur la moyenne prédite (`"pmm"`), régression logistique (`"logreg"`), polytomique (`"polyreg"`) ou à cotes proportionnelles (`"polr"`). Un cycle consiste à imputer successivement $Z_1, \ldots, Z_p$, chacune sachant les valeurs courantes des autres, en tirant d'abord ψ_j^* de sa loi a posteriori approchée. On effectue de 5 à 20 cycles, et on obtient M jeux indépendants (`mice` fait évoluer M chaînes en parallèle).

MICE est un échantillonneur de Gibbs qui converge vers la bonne loi prédictive lorsque les lois conditionnelles sont *compatibles* (engendrées par une loi conjointe) ; ce n'est pas garanti en général, mais la méthode s'avère robuste. Le modèle d'imputation doit être au moins aussi riche que le modèle d'analyse (inclure la réponse, les interactions et les termes non linéaires du modèle d'analyse) : on dit alors que les deux modèles sont *congénères* (Meng, 1994).

8.4 Imputation par augmentation de données

L'augmentation de données (Tanner et Wong, 1987 ; Schafer, 1997 ; Wu, § 3.5) part d'un modèle conjoint, ici $Y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(\mu, \Sigma)$, et construit une chaîne de Markov qui alterne :

Étape I (imputation) tirer $Z_{\text{man}}^{(t+1)} \sim p(Z_{\text{man}} \mid Z_{\text{obs}}, \mu^{(t)}, \Sigma^{(t)})$: pour chaque unité, la loi normale conditionnelle des composantes manquantes sachant les composantes observées ;

Étape P (a posteriori) tirer $(\mu^{(t+1)}, \Sigma^{(t+1)}) \sim p(\mu, \Sigma \mid Z_{\text{obs}}, Z_{\text{man}}^{(t+1)})$, la loi a posteriori en données complétées.

La loi stationnaire de la chaîne est $p(\mu, \Sigma, Z_{\text{man}} \mid Z_{\text{obs}})$; après une période de chauffe, on conserve des jeux complétés espacés pour qu'ils soient approximativement indépendants. Wu (exemple 3.3) présente le cas bivarié sans préciser l'étape P ; la voici.

Lois a priori conjuguées normale–Wishart inverse. On note $\Sigma \sim \mathcal{W}^{-1}(\Psi, \nu)$ la loi de Wishart inverse de densité proportionnelle à $|\Sigma|^{-(\nu+p+1)/2} \exp\{-\operatorname{tr}(\Psi\Sigma^{-1})/2\}$ (équivalemment, Σ^{-1} suit une Wishart de matrice d'échelle Ψ^{-1} et ν degrés de liberté). On prend

$$\mu \mid \Sigma \sim \mathcal{N}_p(\mu_0, \Sigma/\kappa), \qquad \Sigma \sim \mathcal{W}^{-1}(V, \nu_0).$$

(Le document du H26 note κ par τ ; on change de lettre pour éviter la confusion avec les variances résiduelles τ_j^2.) La conjugaison donne, avec $\bar{y}$ et S la moyenne et la matrice de covariance empirique sans biais des données *complétées courantes*,

$$\Sigma \mid y \sim \mathcal{W}^{-1}\Big(V + (n-1)S + \tfrac{n\kappa}{n+\kappa}(\bar{y}-\mu_0)(\bar{y}-\mu_0)^\top,\ \nu_0 + n\Big), \qquad \mu \mid \Sigma, y \sim \mathcal{N}_p\Big(\tfrac{\kappa\mu_0 + n\bar{y}}{\kappa+n},\ \tfrac{\Sigma}{\kappa+n}\Big).$$

Pour une loi a priori peu informative, on prend κ et la diagonale de V petits et positifs, et $\mu_0 = 0$. À la limite $\kappa \to 0$, $V \to 0$, $\nu_0 = -1$, la loi a priori devient $p(\mu, \Sigma) \propto |\Sigma|^{-(p+1)/2}$ (a priori de Jeffreys dit « d'indépendance ») et

$$\Sigma \mid y \sim \mathcal{W}^{-1}((n-1)S,\ n-1), \qquad \mu \mid \Sigma, y \sim \mathcal{N}_p(\bar{y}, \Sigma/n).$$

Étape I dans le cas bivarié. Pour le patron (1,0), tirer $Y_{i2}^* \sim \mathcal{N}\big(\mu_2 + \frac{\sigma_{12}}{\sigma_{11}}(Y_{i1} - \mu_1),\ \sigma_{22} - \frac{\sigma_{12}^2}{\sigma_{11}}\big)$, où (μ, Σ) sont les valeurs tirées à l'étape P ; symétriquement pour le patron (0,1). Attention : le deuxième argument est une *variance* ; dans `rnorm`, il faut passer sa racine carrée.

Le code `aug.donnees` distribué avec les notes implante ces deux étapes ; on tire la Wishart inverse par `solve(rWishart(1, n - 1, solve((n - 1) * S))[, , 1])`.

L'augmentation de données est la méthode « orthodoxe » lorsqu'un modèle conjoint plausible existe (normale multivariée, modèle général de localisation pour variables mixtes, modèles mixtes multivariés avec le module `jomo`). MICE est plus flexible pour des variables de types variés. Avec des lois conditionnelles normales linéaires, les deux méthodes ciblent la même loi prédictive.

8.5 Les règles de Rubin

Soit $\widehat{\theta}^{(m)}$ l'estimation (scalaire ou vectorielle, de dimension k) et $\widehat{V}^{(m)}$ son estimation de variance obtenues sur le m-ième jeu complété. On calcule

$$\bar{\theta} = \frac{1}{M}\sum_{m=1}^{M} \widehat{\theta}^{(m)}, \qquad \bar{W} = \frac{1}{M}\sum_{m=1}^{M} \widehat{V}^{(m)} \quad \text{(variance intra)},$$

$$B = \frac{1}{M-1}\sum_{m=1}^{M} (\widehat{\theta}^{(m)} - \bar{\theta})(\widehat{\theta}^{(m)} - \bar{\theta})^\top \quad \text{(variance inter)}.$$

Définition 8.1 (Règles de Rubin, 1987)**.** L'estimateur combiné est $\bar{\theta}$ et sa variance totale est

$$T = \bar{W} + \Big(1 + \frac{1}{M}\Big)B. \tag{6}$$

Le terme $\bar{W}$ estime le premier terme de (5) (la variance qu'on aurait avec des données complètes), B estime le second (la variance due aux données manquantes), et le facteur $1/M$ tient compte du fait que $\bar{\theta}$ est une moyenne de M tirages seulement : $\mathrm{Var}(\bar{\theta} - \theta_\infty) = B/M$, où θ_∞ est la limite quand $M \to \infty$.

Correction au document du H26

Le document du H26 écrit la variance combinée comme $\frac{1}{M}\sum_m \mathrm{Var}(\widehat{\theta}^{(m)}) + \frac{M+1}{M}(M-1)^{-1}\sum_m (\widehat{\theta}^{(m)} - \bar{\theta})(\widehat{\theta}^{(m)} - \bar{\theta})^\top$. L'expression est correcte *à condition* de comprendre $\mathrm{Var}(\widehat{\theta}^{(m)})$ comme la **variance estimée** $\widehat{V}^{(m)}$ calculée par l'analyse en données complètes du jeu m (par exemple l'erreur type au carré fournie par `lm` ou la matrice sandwich d'une GEE), et non comme la variabilité des $\widehat{\theta}^{(m)}$ entre imputations, qui est l'objet de B. La formule correcte est (6).

Intervalles de confiance et tests. Pour un paramètre scalaire, $(\theta - \bar{\theta})/\sqrt{T}$ suit approximativement une loi de Student. Posons

$$r = \frac{(1 + M^{-1})B}{\bar{W}} \qquad \text{(augmentation relative de la variance)},$$

$$\lambda = \frac{(1 + M^{-1})B}{T} \qquad \text{(proportion de la variance due aux données manquantes)}.$$

Rubin (1987) propose $\nu_{\mathrm{R}} = (M-1)(1+r^{-1})^2 = (M-1)/\lambda^2$ degrés de liberté. Ce nombre peut dépasser les degrés de liberté ν_{com} de l'analyse en données complètes, ce qui est absurde en petit échantillon. Barnard et Rubin (1999) proposent

$$\nu_{\mathrm{BR}} = \left(\frac{1}{\nu_{\mathrm{R}}} + \frac{1}{\nu_{\mathrm{obs}}}\right)^{-1}, \qquad \nu_{\mathrm{obs}} = \frac{\nu_{\mathrm{com}}+1}{\nu_{\mathrm{com}}+3}\,\nu_{\mathrm{com}}(1-\lambda),$$

qui est toujours inférieur à ν_{com} ; c'est ce qu'utilise `pool()` dans `mice` (argument `dfcom`). Pour un paramètre vectoriel, des tests de Wald multivariés combinés existent (statistique D_1 ; Li, Raghunathan et Rubin, 1991), ainsi que des tests du rapport de vraisemblance combinés (D_3).

Choix de M. L'efficacité relative de $\bar{\theta}$ avec M imputations par rapport à $M=\infty$ est $(1+\gamma/M)^{-1}$, où γ est la fraction d'information manquante (estimée par λ ou par la `fmi` de `mice`). Avec $\gamma = 0{,}3$ et $M=5$, l'efficacité est déjà de 94 %, d'où la recommandation historique $M \in \{3,\ldots,10\}$. Cependant, les erreurs types, les degrés de liberté et les valeurs p sont beaucoup plus instables que l'estimation ponctuelle. La règle pratique de White, Royston et Wood (2011) suggère de prendre M au moins égal au pourcentage d'unités incomplètes (par exemple $M=20$ si 20 % des unités ont au moins une valeur manquante) ; avec la puissance informatique actuelle, M entre 20 et 100 est courant.

Niveau doctorat : validité fréquentiste des règles de Rubin

La justification de (6) est bayésienne. Rubin (1987, ch. 4) montre que, si les imputations sont propres et l'analyse en données complètes valide, $\bar{\theta}$ est convergent et T estime correctement (ou de façon conservatrice) sa variance. Le résultat peut échouer si l'imputation et l'analyse ne sont pas congénères (Meng, 1994 : T conservateur si l'imputation est plus riche, $\bar{\theta}$ biaisé si elle est plus pauvre) ou si l'estimateur d'analyse n'est pas efficace, auquel cas T peut sous-estimer la variance (Robins et Wang, 2000, proposent un estimateur toujours convergent).

8.6 Exemple en R : imputation « à la main » et avec `mice`

On reprend le jeu de données MAR de la section 5. La fonction `rubin` implante (6) pour μ_2, avec $\widehat{V}^{(m)} = s_2^{2(m)}/n$.

```
rubin <- function(est, v) {
  M <- length(est); qb <- mean(est); W <- mean(v); B <- var(est)
  Tt <- W + (1 + 1/M) * B; r <- (1 + 1/M) * B / W
  c(est = qb, W = W, B = B, T = Tt, se = sqrt(Tt), dl = (M - 1) * (1 + 1/r)^2,
    lambda = (1 + 1/M) * B / Tt)
}
imp1 <- mice.main(Y, M = 20)              # équations chaînées, algorithme ci-dessus
rubin(sapply(imp1, function(D) mean(D[, 2])), sapply(imp1, function(D) var(D[, 2]) /
    200))
imp2 <- aug.donnees(Y, M = 20)            # augmentation de données, Jeffreys
rubin(sapply(imp2, function(D) mean(D[, 2])), sapply(imp2, function(D) var(D[, 2]) /
    200))
```

```
     est        W        B        T       se       dl   lambda     # MICE à la main
  1.9202   0.0041   0.0011   0.0053   0.0726 397.2961   0.2187
     est        W        B        T       se       dl   lambda     # augmentation de données
  1.9072   0.0041   0.0009   0.0051   0.0711 493.4732   0.1962
```

Avec `mice`, l'analyse se fait par `with()` et la combinaison par `pool()` :

```
library(mice)
D <- as.data.frame(Y)
imp <- mice(D, m = 20, method = "norm", maxit = 10, seed = 7360, printFlag = FALSE)
fit <- with(imp, lm(Y2 ~ 1))            # mu2 : ordonnée à l'origine
pool(fit)
```

```
Class: mipo    m = 20
         term  m estimate        ubar           b           t dfcom       df
1 (Intercept) 20 1.922554 0.004097955 0.001097589 0.005250423   199 110.6375
        riv    lambda       fmi
1 0.2812301 0.2195001 0.2332367
```

Dans la sortie de `pool`, `ubar` est $\bar{W}$, `b` est B, `t` est T, `riv` est r et `df` est ν_{BR} (ici 110,6, contre $\nu_{\mathrm{R}} \approx 394$ avec la formule de Rubin de 1987). L'erreur type combinée (0,072) est très proche de celle de l'EMV (0,071) et nettement plus grande que l'erreur type naïve de l'imputation simple (0,058). Pour un modèle de régression, on procède de même (`with(imp, lm(Y2 ~ Y1))`) et `summary(pool(...), conf.int = TRUE)` donne les intervalles combinés. Environ 22 % de la variance de $\bar{\mu}_2$ est attribuable aux données manquantes.

9 Étude de simulation : MCAR, MAR et MNAR

Pour comparer les méthodes, on génère $B = 300$ jeux de $n = 200$ observations sous chacun des trois mécanismes de l'exemple 2.3 (environ 35 % d'unités incomplètes), et on estime $\mu_2 = 2$ et $\rho = 0{,}5$ par : les données disponibles ; l'EMV (EM, erreur type par la hessienne observée) ; l'imputation simple par régression avec erreur type naïve ; l'imputation multiple (`mice`, méthode `"norm"`, $M = 10$, règles de Rubin). Ce code reprend et simplifie le fichier `simu.rtf` du H26.

Table 1 – Simulation (300 réplications, $n = 200$) : biais, écart type empirique, moyenne des erreurs types estimées et couverture de l'intervalle à 95 % pour $\mu_2 = 2$; biais pour $\rho = 0{,}5$.

Mécanisme	Méthode	Biais (μ_2)	ET emp.	ET moy.	Couverture	Biais (ρ)
MCAR	Données disponibles	−0,001	0,079	0,078	0,943	−0,003
	EMV	0,000	0,077	0,076	0,950	−0,002
	Imputation simple	0,001	0,077	0,066	0,897	0,074
	IM (`mice`)	−0,001	0,078	0,077	0,953	−0,006
MAR	Données disponibles	−0,098	0,079	0,078	0,740	−0,109
	EMV	−0,006	0,078	0,078	0,960	−0,003
	Imputation simple	−0,066	0,077	0,063	0,747	−0,032
	IM (`mice`)	−0,006	0,078	0,079	0,960	−0,007
MNAR	Données disponibles	−0,184	0,076	0,076	0,340	−0,110
	EMV	−0,127	0,076	0,076	0,590	−0,043
	Imputation simple	−0,164	0,072	0,062	0,307	−0,020
	IM (`mice`)	−0,128	0,076	0,077	0,597	−0,048

Les résultats (tableau 1 et figure 2) confirment la théorie.

— **MCAR** : toutes les méthodes sont essentiellement sans biais pour μ_2. L'EMV et l'imputation multiple sont à peine plus précises que les données disponibles (avec $\rho = 0{,}5$, l'autre variable

apporte peu d'information sur μ_2). L'imputation simple a une couverture insuffisante (90 %) : son erreur type naïve est trop petite (proposition 5.1), et elle surestime ρ.

— **MAR** : les données disponibles sont biaisées et leur couverture tombe à 74 % ; l'EMV et l'imputation multiple restent essentiellement sans biais avec une couverture proche de 95 %. L'imputation simple ne corrige qu'un tiers du biais de la moyenne : la régression d'imputation est estimée sur les cas complets, dont la sélection dépend ici aussi de la variable à imputer (proposition 3.3 : la sélection sur la réponse biaise la régression). Elle sous-estime en outre l'incertitude.

— **MNAR** : aucune méthode n'est sans biais. Les méthodes fondées sur MAR réduisent le biais (l'information portée par l'autre variable aide), sans l'éliminer.

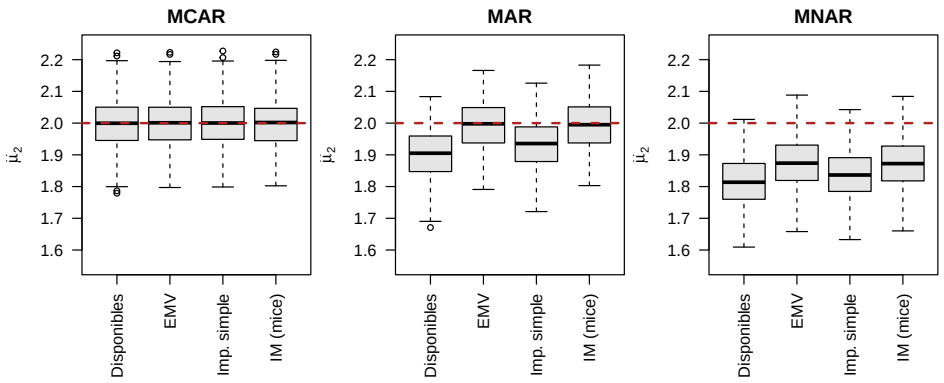


FIGURE 2 – Distribution de $\widehat{\mu}_2$ sur 300 réplications pour chaque mécanisme et chaque méthode ; la ligne pointillée indique la vraie valeur $\mu_2 = 2$.

10 Synthèse

Grille méthodologique

1. **Objet d'intérêt** : le paramètre θ du modèle des données complètes $p(z \mid \theta)$ (moyennes, coefficients de régression, etc.).
2. **Nuisance, données manquantes, biais** : les valeurs Z_{man}, le mécanisme $p(r \mid z, \psi)$ et, avec des covariables manquantes, la loi des covariables. L'analyse des cas complets est biaisée hors MCAR.
3. **Approche** : sous MAR et paramètres distincts, ignorer le mécanisme (théorème 4.2) et intégrer (vraisemblance) ou simuler (imputation multiple) les valeurs manquantes à partir d'un modèle pour les données. Sous MNAR, analyse de sensibilité.
4. **Critère** : vraisemblance des données observées ; ou, pour l'imputation multiple, le critère de l'analyse en données complètes appliqué à chaque jeu imputé.
5. **Calcul** : Newton-Raphson ou EM ; tirages a posteriori (équations chaînées, augmentation de données par MCMC).
6. **Inférence** : information observée ; règles de Rubin ($T = \bar{W} + (1 + M^{-1})B$, degrés de

liberté de Barnard-Rubin).

À retenir

- Le mécanisme se définit par la loi de R sachant Z : MCAR (indépendance), MAR (dépendance aux seules données observées), MNAR (dépendance aux valeurs manquantes). On ne peut pas distinguer MAR de MNAR avec les données observées.
- Les cas complets ne sont sans biais que sous MCAR ; le biais limite pour une moyenne est $\mathrm{Cov}\{Y, \pi\}/\mathrm{E}(\pi)$. Exception : en régression, une non-réponse qui ne dépend que des covariables ne biaise pas les coefficients.
- Sous MAR et paramètres distincts, la vraisemblance des données observées suffit (ignorabilité). Les covariables manquantes exigent un modèle pour leur loi.
- L'imputation simple sous-estime la variance. L'imputation multiple propre tire les paramètres du modèle d'imputation de leur loi a posteriori, puis les valeurs manquantes ; on combine par $T = \bar{W} + (1 + 1/M)B$.
- MICE spécifie des lois conditionnelles variable par variable ; l'augmentation de données part d'un modèle conjoint (normale–Wishart inverse).
- Sous MNAR, aucune méthode n'est convergente sans hypothèse non vérifiable : on fait une analyse de sensibilité (modèles de sélection, de mélange de patrons).

11 Exercices

Exercice 11.1 (Exercice d'imputation, H24–H26, question 1)**.** Considérons deux variables V et Y dont seule Y a des valeurs manquantes, et supposons que la probabilité qu'elle manque dépende de la vraie valeur de Y. On veut estimer $\mathrm{E}[V \mid Y]$, l'espérance conditionnelle de V sachant Y, par exemple en supposant une relation linéaire entre V et Y et en estimant une droite de régression.

a) De quel type de données manquantes s'agit-il (MCAR, MAR ou MNAR) ?

b) L'analyse des cas complets est-elle biaisée dans ce contexte ? Étudiez le ratio $\mathrm{E}[VR \mid Y]/\mathrm{E}[R \mid Y]$, sans préciser de forme pour l'estimateur de $\mathrm{E}[VR \mid Y]$. Comparez avec la proposition 3.3.

Exercice 11.2 (Exercice d'imputation, H24–H26, question 2 : croissance dentaire)**.** Le fichier `Dev1Ex1.txt` contient la distance entre deux points de la mâchoire, mesurée sur radiographie chez 27 enfants aux âges de 8, 10, 12 et 14 ans (une ligne par enfant ; 11 filles, puis 16 garçons). Des valeurs manquantes ont été créées artificiellement selon un mécanisme MAR (`dentaire_manquant.txt`). On ajuste le modèle `y ~ age * sexe` par GEE avec corrélation de travail échangeable. Traitez les données manquantes par l'analyse des cas complets (`na.action = na.omit` dans `gee`, ou retrait préalable des réponses manquantes pour `geeglm`) et par imputation multiple avec `mice`, et comparez aux données pleines. Examinez d'abord le patron : est-il monotone ? Quelles variables prédisent la non-réponse ? (Repris à la séance 9.)

Exercice 11.3 (Examen final H23, question 3)**.** On reprend l'étude de Giroux et coll. (2012) de l'exemple 1.1, où des modèles de régression linéaire multiple relient la DMO au génotype UGT2B17 (double délétion contre autres génotypes) et à l'hormonothérapie, en ajustant pour d'autres facteurs. Le tabagisme (jamais, ancienne, actuelle) manque chez 242 femmes, l'activité physique (moins d'une,

une à deux, trois activités ou plus par semaine) chez 26 femmes et le génotype (2 allèles, 1 allèle et 1 délétion, 2 délétions) chez 4 femmes ; la DMO, l'âge et le poids sont toujours observés.

Proposez et justifiez une méthode pour traiter les valeurs manquantes dans cette analyse. Écrivez le modèle de régression ou de prédiction requis par la méthode proposée pour une des variables ayant des valeurs manquantes, s'il y a lieu.

Exercice 11.4 (Questions de pratique de l'examen final, H23, question 1 a)–b))**.** Des enquêtes de couverture vaccinale (statut complet oui/non) sont menées au Québec auprès d'un nombre égal d'enfants de 1 an et de 2 ans ; on suppose la non-réponse complètement aléatoire. Le statut est évalué à 1 an (Y_1) et à 2 ans (Y_2) : seul Y_1 est observé chez les enfants de 1 an. Des prédicteurs sociodémographiques X (âge, scolarité de la mère, etc.), observables à 1 an, sont toujours mesurés. On veut estimer un modèle log-linéaire de $\mathbb{P}(Y_2 = 1 \mid X)$ par les équations $\sum_i X_i\{y_{2i} - \exp(X_i^\top \beta)\} = 0$ (si les données pleines étaient disponibles).

a) Identifiez l'objet statistique d'intérêt et les paramètres de nuisance de dimension infinie dans l'analyse semi-paramétrique de cette enquête.

b) Proposez un modèle de régression pour imputer les valeurs manquantes de Y_2. Quel est le mécanisme de données manquantes pour Y_2 ?

Références

— Barnard, J. et Rubin, D. B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika*, 86, 948–955.

— Dempster, A. P., Laird, N. M. et Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *JRSS B*, 39, 1–38.

— Diggle, P. et Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Applied Statistics*, 43, 49–93.

— Giroux, S. et coll. (2012). UGT2B17 gene deletion associated with an increase in bone mineral density [...]. *Osteoporosis International*, 23, 1163–1170.

— Kenward, M. G. et Molenberghs, G. (1998). Likelihood based frequentist inference when data are missing at random. *Statistical Science*, 13, 236–247.

— Little, R. J. A. et Rubin, D. B. (2019). *Statistical Analysis with Missing Data*, 3e éd. Wiley.

— Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 9, 538–558.

— Robins, J. M. et Wang, N. (2000). Inference for imputation estimators. *Biometrika*, 87, 113–124.

— Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63, 581–592.

— Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley.

— Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall.

— van Buuren, S. (2018). *Flexible Imputation of Missing Data*, 2e éd. Chapman & Hall/CRC.

— White, I. R., Royston, P. et Wood, A. M. (2011). Multiple imputation using chained equations : issues and guidance for practice. *Statistics in Medicine*, 30, 377–399.

— Wu, L. (2010). *Mixed Effects Models for Complex Data*, ch. 3. Chapman & Hall/CRC.